DOCUMENT RESUME

ED 084 306 TM 003 329

AUTHOR Fennessey, James
TITLE Using Achievement Growth to Analyze Educational Programs. Work Unit 2A, Reward Structures-Achievement Growth.
INSTITUTION Johns Hopkins Univ., Baltimore, Md. Center for the Study of Social Organization of Schools.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
REPORT NO R-151
PUB DATE Mar 73
CONTRACT NE-C-00-3-0114
NOTE 39p.

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Academic Achievement; *Achievement Gains; *Achievement Tests; Decision Making; *Educational Accountability; Grade Equivalent Scores; Measurement; *Program Evaluation; Scores; Standardized Tests; Student Improvement; Test Results
IDENTIFIERS Interval Scales; *Scaling

ABSTRACT

The single most important output of any school is probably the magnitude of its students' growth in academic achievement. A variety of standardized tests has been developed to measure aspects of this achievement; however, only recently have administrators attempted to use such tests to help review and make decisions about educational programs. There have been such applications of achievement tests recently, as well as associated problems. One often unrecognized problem is that for these program analysis applications, it is necessary to develop a score format appropriate to the decision context, and one which has the properties of an interval scale. There have been some difficulties inherent in past attempts to develop internal scales of academic achievement; these difficulties carry several implications. With a more open-minded and progmatic approach research and development work on some of these issues can be done rather easily and inexpensively. (Author/NE)

Center for Social Organization of Schools

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED 084306

Report No. 151

March, 1973

USING ACHIEVEMENT GROWTH TO ANALYZE EDUCATIONAL PROGRAMS

James Fennessey

EM 003 368

The Johns Hopkins University

FILMED FROM BEST AVAILABLE COPY

STAFF

John L. Holland, Director

James M. McPartland, Assistant Director

Joan E. Brown

Judith P. Clark

David L. DeVries

Keith J. Edwards

Gail M. Fennessey

James J. Fennessey

Stephanie G. Freeman

Ellen Greenberger

Edward J. Harsch

Samuel T. Helms

John H. Hollifield

Patricia A. Hughes

Nancy L. Karweit

Samuel A. Livingston

Edward McDill

Alyce J. Nafziger

Dean H. Nafziger

Karen A. Schwartzman

John P. Snyder

Julian C. Stanley

Gerald D. Williams

ED 084306

USING ACHIEVEMENT GROWTH TO ANALYZE EDUCATIONAL PROGRAMS

CONTRACT NO. NE-C-00-3-0114

WORK UNIT 2A

REWARD STRUCTURES-ACHIEVEMENT GROWTH

JAMES FENNESSEY

REPORT NO. 151

March, 1973

Published by the Center for Social Organization of Schools, supported in part as a research and development center by funds from the United States National Institute of Education, Department of Health, Education, and Welfare. The opinions expressed in this publication do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the Institute should be inferred.

The Johns Hopkins University

Baltimore, Maryland

INTRODUCTORY STATEMENT

The Center for Social Organization of Schools has two primary objectives: to develop a scientific knowledge of how schools affect their students, and to use this knowledge to develop better school practices and organization.

The Center works through five programs to achieve its objectives. The Academic Games program has developed simulation games for use in the classroom. It is evaluating the effects of games on student learning and studying how games can improve interpersonal relations in the schools. The Social Accounts program is examining how a student's education affects his actual occupational attainment, and how education results in different vocational outcomes for blacks and whites. The Schools and Maturity program is studying the effects of educational experience on a wide range of human talents, competencies, and personal dispositions in order to formulate -- and research -- important educational goals other than traditional academic achievement. The School Organization program is currently concerned with authority-control structures, task structures, reward systems, and peer group processes in schools. The Careers and Curricula program bases its work upon a theory of career development. It has developed a self-administered vocational guidance device and a self-directed career program to promote vocational development and to foster satisfying curricular decisions for high school, college, and adult populations.

This report, prepared by the School Organization Program, discusses the applications of output measures in the operation and improvement of schools, examines some difficulties in obtaining satisfactory measures of achievement growth, and outlines a new approach for developing achievement scores to analyze educational programs.

ABSTRACT

The single most important output of any school is probably the magnitude of its students' growth in academic achievement. A variety of standardized tests have been developed to measure aspects of this achievement; however, only recently have administrators attempted to use such tests to help review and make decisions about educational programs. This paper describes some examples of these recent applications of achievement tests and discusses some of the associated problems. In particular, one often unrecognized problem is noted: for these program analysis applications, it is necessary to develop a score format appropriate to the decision context, and one which has the properties of an interval scale. Some difficulties inherent in past attempts to develop interval scales of academic achievement are described, and several implications of these difficulties are mentioned. Finally, the suggestion is made that, with a more open-minded and pragmatic approach, research and development work on some of these issues can be done rather easily and inexpensively. Such an approach is outlined in the concluding section of the paper.

ACKNOWLEDGMENT

The author acknowledges with appreciation the helpful comments of Gail M. Fennessey, Nancy L. Karweit, James M. McPartland and I. Richard Savage on an earlier draft.

This report is based in part on a paper prepared for presentation at the 1972 Annual Meeting of the American Sociological Association, New Orleans, Louisiana, 1972.

INTRODUCTION

Adequate measurement of the outputs of schools, especially academic achievement in reading and arithmetic, is a logically necessary condition for evaluation research. In addition, however, from a school organization point of view, such measurement is a prerequisite if rewards in the school are to be distributed responsively and so be instrumental in mobilizing motivation. It is also needed if school personnel are to make intelligent operating decisions about allocation of resources and to plan for future demands.

The first section of this report reviews some of the recent developments in educational practice and theory which have caused new interest to be focused upon output measures. It also indicates that, as a consequence of this recent attention, a number of new problems with the tests used to measure academic achievement have come to light. The second section discusses some of the history behind these problems and the limited applicability of previous work in helping to produce solutions to them. In particular, it will be argued that the primary criteria for a successful scale are that (1) the scale should measure exactly the variable that the user is concerned with in his decision-making process; and (2) the scale should have interval properties that are adequately justified. In most of the discussions now in the literature which deal with the use of achievement tests in educational program analysis, both of these criteria have been perceived only in an approximate and vague way. The third section of the paper outlines a new approach to the development and use of scales of achievement <u>growth</u> for use in educational program

analysis decisions. It is argued that this approach is conceptually defensible and practically feasible. Although the new approach is a straightforward one, it leads to some surprising inferences and illuminates some neglected issues.[1]

This whole topic of achievement tests and their use bristles with a number of controversial issues. The present paper will not address itself to most of these, even though it is recognized that they are important and closely related to the issues upon which it does focus. Among the related issues which cannot be directly discussed here are: the narrowness of content coverage of achievement tests, the degree of cultural bias of achievement tests, the validity of these tests in various circumstances, the reliability of the tests, the statistical analysis of change scores (see, for example, Harris, 1963; Werts and Linn, 1970; Cronbach and Furby, 1970), or the use of practical tasks, criterion-referenced tests, and course grades as output measures.

---

[1]The argument of this paper could have been presented differently. The situation could have been analyzed by stating that the task confronting the educational measurement expert is to find ways to specify a utility function for different amounts of educational change. See Nelvin Lifson (1972) for a clear discussion of this approach. Lifson indicates that there are two approaches in general use for developing utility functions. The first is the "standard gamble" approach, following the work of Von Neumann and Morgenstern. The second is the direct magnitude estimation method developed and used by S. S. Stevens. However, the approach we suggest in this paper is different from either of these, and was selected to allow us to move from familiar to less familiar ideas. A future report will present an explicit comparison of these three approaches to the development of a utility scale of achievement outcomes.

THE RECENT INTEREST IN PERFORMANCE MEASURES

Previous work (Cohen and Filipczak, 1971; Kirschenbaum, Napier, and Simon, 1971) suggests that the ways in which schools as organizations monitor and reward their participants are, in general, clumsy and unfair. As a result, schools do not fully utilize an important potential motivational force. Others (e.g., Owens, 1970) charge that school administration tends to be either too mechanized (blindly following procedures because they exist); or else too intuitive (following personal hunches, and making decisions on the basis of inarticulate feelings).

A number of causal factors operate to inhibit the development of responsive and rational school administration. One is the lack of consensus about the relative importance of the different objectives a school might adopt (Stake, 1970). In the absence of such consensus, procedures and priorities for achieving the objectives cannot be rationally established. Another obstacle is the need to discover and use motivational inducements for the staff and students which are effective and feasible. However, even if these two sets of problems are dealt with, additional problems arise in creating and implementing procedures for actually collecting the information on which to base decisions, and for actually delivering responses as intended.

These problems vary in form, depending upon the particular school setting and the particular output objective under consideration. At first glance, it would seem that many of these problems would be least serious for outputs such as reading achievement or mathematics achievement. For these outputs, there is a general consensus that they are highly important. Also, because information about these outputs appears on school transcripts,

it influences the kinds of subsequent opportunities that will be offered to a child. In other words, there are at least some motivational inducements structurally attached to the academic achievement of a student. Procedurally, data on these academic achievements are collected in two ways -- by teacher-determined grades, and by standardized tests.

Few claims are made in the research literature for the accuracy, precision, or clarity of teacher-determined grades (Warren, 1971; Donaldson, 1971). Traditionally, grades have been regarded as crude indicators. In practice, however, they have been generally accepted as serviceable (particularly when several grades for an individual student are averaged) as a means of providing an approximate measure of that student's level of competence.

Standardized achievement tests are less widely used than grades; however, they are quite commonly administered, and the trend in this direction seems to be increasing. It is generally felt that standardized tests are comparable across a wider variety of school situations. Also, they provide a more precise and objective indication of a student's level of knowledge than teacher-determined grades.

However, these standardized tests are subject to a large number of limitations. These limitations, as well as the generally undeveloped state of the art in the measurement of academic output, have become apparent recently as a by-product of several large scale efforts to evaluate educational programs. Most of these evaluation efforts have been sponsored by the Federal government; they include the Equality of Educational Opportunity report (1966), the evaluation of Project Headstart (1969), and the report of the experiment in performance contracting (1972).

These evaluation-research efforts (and numerous other recent studies as well) used standardized tests to measure the school's output. Previously, achievement tests had been used for program review only in much smaller projects and in the context of purely academic research. The large size, the high public interest, and the political sensitivity of these recent efforts all have contributed to the generation of considerable controversy about their findings and their methods (see, for example, Bowles and Levin, 1968; Cain and Watts, 1970; Campbell and Erlebacher, 1970; Smith and Bissell, 1970; Guthrie, 1970). As a result, it has been concluded that there are a number of difficulties involved in this kind of use of standardized achievement tests, (e.g., Stake, 1967; Popham, 1972; Campbell and Erlebacher, 1970; Klein, 1971; Coleman and Karweit, 1972). Under these circumstances, any improvements that can be made to reduce these problems will benefit future research aimed at the large-scale evaluation of educational programs.

Obviously, the difficulties faced by researchers and by educational administrators are not identical. Various writers (Cohen, 1970; Fennessey, 1972; Rossi and Williams, 1972) have enumerated some of the difficulties confronting the evaluation researcher. In some instances, these distinctive problems have been described by the researchers themselves (e.g., Planar Corporation, 1972). Various other writers have outlined some of the problems educational administrators face when they consider the use of output measures. However, the issues faced by each of these groups are similar enough that a set of procedures aimed to benefit one would also considerably help the other.

The utility of standardized tests was regarded as limited even prior to their recent use to evaluate educational programs. More recently, a number of writers (Dyer, 1971; Lennon, 1971; Rivlin, 1971) have lamented the extreme emphasis placed upon such tests as the criterion variable. Some writers have objected that the usual standardized tests encompass too narrow a domain (Nash and Agne, 1972). In other words, they see the tests as inadequate indicators of the outcomes and objectives of an educational program.

More technically, there has been considerable debate over the appropriateness of the different available score formats for measuring academic achievement as part of the quantitative analysis of an educational program. The grade-equivalent score format in particular has been severely criticized (Dyer, 1971; Coleman and Karweit, 1972), primarily because of a property that has been called "fan-spread" (Campbell, 1971). The issues in this debate have been many and complex. Its content has been highly technical, and its tone in many cases highly polemic. Yet, almost all the writers have neglected to consider some really fundamental points about score uses and score format. This paper suggests two of these fundamental points. It then indicates their implications for the future use of standardized achievement tests in connection with decisions about educational programs, whether these are operating decisions or research decisions.

THE MEASUREMENT OF DIFFERENCES IN ACADEMIC PERFORMANCE

Review of Current Practice in Score Construction

The first basic, but often neglected, point to be made about score format is that the development of an appropriate score format is possible only if the decision context for which the measurement is to be used is made explicit. The second basic point is that such development is accomplished only when the resulting score can be shown to be an interval scale with respect to the quantity being measured and the comparisons being made. In less formal words, one must be sure that he is measuring the correct variable for his purposes and that the measure is strong enough to support the arithmetic operations required by the usual analysis techniques (i.e., parametric statistics).

That the choice of score format depends upon the user's purpose is a point sometimes made in the materials accompanying the currently published tests, but this caution is directed only to the context of evaluating the score of an individual student. It has not been generally recognized that the context of comparing and evaluting educational programs imposes quite a different set of demands than does using the scores to locate an individual student.

The second point, that an interval scale is necessary, has gone equally unrecognized. The simple fact is that if the scores being used do not form an interval scale with respect to the trait being measured, then it makes no sense to add them together, or to perform the other operations of ordinary arithmetic. Without the use of such operations, almost all the techniques customarily used for analysis cannot be employed

(see Miller and Starr, 1967).

The invisibility of this requirement for interval scores probably arises because it is mechanically possible to add scores together and compute, say, an average score. The thrust of the point here, however, is that if the scale being used is not an interval scale of the intended trait, then the results of such operations will not mean substantively what they appear to mean. That is, although two classrooms might have identical average achievement scores, the true level of achievement in one class might be quite different from the true level of achievement in the other class. Conversely, two classes might have average achievement scores that are quite different, yet their true levels of achievement might be nearly the same. This phenomenon would be severe to the extent that the scale being used were not a linear transformation of the true scale over the range of scores encountered in the two classrooms. Coleman and Karweit (1972) discuss a related point, namely, the implications of using scales that are not related to each other linearly. Their cautions apply equally to any scale that is not linearly related to the desired underlying trait.

A second reason that the necessity for establishing an interval scale is little understood by most users of achievement test data is that the publishers of these tests have concentrated primarily on developing and providing score forms that are (1) simple and intuitively meaningful, and (2) appropriate for describing the relative academic achievement of an individual child. Until recently, there has been no corresponding demand for the development of scores that would be appropriate for other uses. In the absence of such demand, publishers understandably have been

more or less silent about this limitation in the applicability of their available scores.

A quick survey of some of the literature on achievement testing (S.R.A., 1969; Lindquist and Hieronymus, 1964; Coleman and Karweit, 1972) reveals that there is no consensus in the field about (1) how to measure changes (or "growth") over time in academic performance; or (2) how to compare the academic performance of two groups of students. Only in some fairly obscure technical publications are these two problems seen to be basically identical (e.g., E. F. Gardner, 1947). Each can be reduced simply to a demand for a score format that has the properties of an interval scale. Strangely enough, little of the available literature discusses the real problems involved in establishing interval scales for academic achievement. McNemar (1942) discusses related problems for the Stanford-Binet intelligence test; but his work is concerned only with that particular test, used as a measure of learning potential. The few pieces of work that have been done (e.g., Flanagan, 1939) are cited in many places, but have not been extended or updated.

To indicate that the issue is one on which there is little agreement, it can be noted that three of the most popular tests use three different approaches to deal with this set of issues. The Iowa Test of Basic Skills is one of the most widely used series of achievement tests. The tests were developed and the manuals written by a well-known psychometrician, E. F. Lindquist. The I. T. B. S. manual, (Lindquist and Hieronymus, 1964, page 14) recommends the use of grade-equivalents for measuring growth. However, the new Metropolitan Achievement Tests series, another widely used battery, offers a special score called "standard

scores" and recommends these standard scores as appropriate for measuring growth. Yet, the M. A. T. manuals give little information about the derivation of these scores and do not say why they are appropriate for indicating growth. A third approach, preferable at least because it is more explicit, is exemplified by the Science Research Associates Achievement Series. This publisher has developed a set of "growth scores" and has prepared a special manual to explain these scores. This manual is both readable and thorough, which is no small achievement in itself. Upon investigation, it seems (cf. Orr, 1972) that the S. R. A. growth scores are derived by essentially the same procedure used in the derivation of the Metropolitan Standard Scores. One irritating aspect of this situation is that the details of the Metropolitan scores derivation procedure are not described by the publisher in any available written form.

When one publisher recommends grade-equivalents for use in measuring growth, another rejects grade-equivalents and provides a special but obscure score for growth, and a third prepares a special set of scores and devotes a lengthy manual to discussing them, it seems clear that there is considerable disagreement in the trade about the correct procedures for justifying the claim that some set of scores has the desired properties. The most interesting point about this situation, however, is that the discussion which occurs does not deal with the procedures that might be used to develop an adequate scale, but instead merely repeats the exhortation to use one or another of the existing score forms.

Constructing Interval Scales for Academic Achievement

The key property of an interval scale is that units at any one point on the scale are the same in real and relevant magnitude as units at any other point on the scale. The important question in this instance is: how can we create a scale for academic output that can be shown to have these interval properties?

In creating an interval scale for academic achievement, there is no way to establish directly that any given proposed scale has the desired properties. That is, the most obvious way to show that a scale is indeed an interval scale is to find a physical operation that corresponds to addition, and a physical relationship that corresponds to equality, and then to show that units which are numerically equal on the scale are indeed physically equal, and that combinations (sums or differences) of units which are numerically equal are in fact physically equal also (Coleman, 1964). This direct justification of an interval scale can be done quite easily with some physical quantities such as length or weight. It cannot be done at all, however, for academic achievement, since for this variable there is no physical operation that is physically the same as adding the two amounts, nor is there any physical relationship corresponding to equality.

There is a second possible approach for establishing that a given scale has interval properties, and it is used in physical measurement as well as in the social sciences. The approach is to adopt a premise, based upon substantive reasons, that the underlying trait in question (in this case, academic achievement) has a certain specified frequency distribution in a certain specified population. Using actual raw scores

from a representative sample of this population, the premise is then exploited by carrying out transformations of the initial (e.g. raw scores) scale until the distribution of scores on the resulting transformed scale has the same shape as that postulated for the underlying trait. This approach is basically an instance of construct validation, but it is not the content of the scale that is being validated, but rather its metric with respect to the trait.

This distributional approach has been applied successfully to the creation of scales for the measurement of intelligence. The premise has been adopted, based upon substantive reasoning about the factors which determine true intelligence, that in a large, unselected population of normal persons, the distribution of the trait "intelligence" will be approximately "Gaussian" or normal. If this initial premise is defensible, then scales derived by using it are also defensible. The intelligence tests most widely used at present have scale scores developed using this line of reasoning, which was first suggested by Thurstone (1925) and later refined by Flanagan (1951) and others.

This same kind of reasoning -- beginning with an assumed shape for the distribution of the true trait in a specified set of persons -- has been used by some publishers (e.g., S.R.A.) of achievement tests in their efforts to create interval scores. The difficulty with using this approach for academic achievement is that there is no compelling reason to assume that the distribution of achievement scores has any particular shape, much less that it is normal. No matter what population is chosen for study, there is bound to be some diversity of educational experience that will affect the shape of the resulting distribution of true scores

on academic achievement. In other words, the argument that the trait is determined by a very large number of statistically independent, individually small causes does not apply nearly as well to academic achievement as it does to intelligence. One writer who has recognized this, at least partially, and attempted to deal with it, is Eric Gardner. Gardner (1950, 1947) advocated the use of a more relaxed distributional assumption than normality. His suggestion is that a distribution which allows for skewness as well as having a general bell-shape (namely, the Pearson Type III) be used. Gardner developed a procedure for creating scale scores based upon this distribution. However, while this procedure does remove one aspect of the restrictiveness of the normal curve assumption, it does not answer the basic objection -- namely, that it is unwarranted to posit any particular distributional form.

There is no intent here to claim that the achievement scales now in use are completely unreasonable. On the contrary, it is likely that they do reflect the ordinal relations between achievement levels perfectly. Moreover, they probably are not extremely distorted, particularly over short ranges, from a true interval scale. However, in many situations where programs are being evaluated, the distortion might be large enough that the difference in scales would make a difference in the final result. In other words, although the distortion may be small, even small distortions could cloud the basic issue as to the relative effectiveness of two programs.

The more general point about the construction of achievement scales is that such scales ought to be chosen to fit the purpose of the decision maker who will use them. According to this criterion, several of the

usual scales, including the grade-equivalent scale, are appropriate for use by educational planners and counselors when they attempt to choose material for curriculum units or make placement decisions about individual students.

For example, many curriculum packages are designed for students whose achievement levels fall within a certain range. If a student's score is outside this range, then another package would be more appropriate for him, so the decision is simply which curriculum to use with a given child. The grade-equivalent scale, regardless of any interval properties, is frequently and correctly used in making such decisions. In other words, the grade-equivalent score format is well-suited for matching an individual student with the curriculum material.

In a different class of situations, the percentile score can be quite useful, also without regard to any interval properties it may or may not possess. Many educational decisions involve a competitive admissions process. In these kinds of decisions, a finite number of places are available, and there are more applicants than can be accommodated. To fill the places, the students whose scores are highest are chosen in order of score until all the available places are filled. For this kind of decision, only the ordinal properties of the scale are needed. The percentile score presents this ordinal information in a convenient and general way, thus simplifying the task of the decision maker and the applicant.

Working from a somewhat different starting point, there has recently been a growing movement among testing experts and educators toward what are referred to as criterion-referenced tests. The scoring associated

with such tests is designed to relate the child's actual achievement to a set of real-life tasks. In other words, there is no claim that this type of test is particularly useful for the evaluation of educational programs, but instead that it locates students directly on dimensions that have clear meaning and interest for parents and prospective employers. These criterion-referenced tests, in other words, link an individual student's skill level to some common real-world situations.

For the person who needs to compare two educational programs, however, these scales provide little help. There is no claim that these score formats provide an interval scale in the context in which they are used, but only an ordinal one. Each is aimed primarily at being useful in decisions about individual students, whether for further schoolwork or for life-work.

This, then, describes the current situation with regard to achievement test score formats and the analysis of educational programs. Before suggesting an alternative approach that seems to show some promise, we need to examine briefly some implications of the existing state of affairs.

Aside from the ambiguities and controversies that inappropriate achievement scales create in efforts to evaluate particular educational programs, an additional consequence is that the confusion and possible distortion in the scales has aggravated the controversy about various approaches to the education of low-income children. A variety of topics relating to racial differences between blacks and whites in learning rates also are latent in these debates. In fact, it seems quite probable that these racial issues have motivated far more discussions of testing

than would be apparent on the surface. These discussions, already tense, frequently become more heated and circular because of the confusion about test scores and their appropriate use in comparative assessments.

A final implication is that the lack of consensus among testing experts has been one major force retarding the introducution of general outcome monitoring (cf. Blau and Scott, 1962) and accountability programs on a routine basis in school operations. The reluctance of teachers and administrators to let themselves be measured by a possibly biased instrument is understandable. What is less understandable is the reluctance of many experts to attack this set of technical issues forthrightly and empirically.

A DECISION-THEORETIC APPROACH

The preceding sections of this paper have indicated that the present use of achievement test scores in program review and evaluation is essentially chaotic. It has been argued here that this chaos arises because insufficient attention has been paid to the logical requirements a set of scores must possess if they are to be useful for a particular purpose. Perhaps the very sophistication of achievement test development and norming procedures has made them apparently unassailable, and so in turn has made these other requirements less detectable.

The Objectives of the Approach

Our point is that, for program review decisions, it is necessary to measure specifically the program's impact on the child, not the actual level of knowledge of the student. The analyst is interested not in

achievement level, but in change of level. The scale he needs is a scale of changes, or growth, not a scale of levels. Recognizing this fact, we see that the scales offered by S.R.A. and the M.A.T. are irrelevant for comparing programs. They have interval properties, alledgedly, in terms of the actual amount of some knowledge that a student possesses, but this is not the concern of program evaluation.

A second point is that the scores used should have interval properties regardless of where growth occurs on the knowledge curve. That is, the scales should have interval properties when comparisons are made between two students who start at different initial levels, or between a student's growth during time interval 1 and his growth during time interval 2. The question is how such scores can be developed without basing them on the distributional form of construct validation. The alternate approach we propose is that of calibrating the new scale for achievement growth against a known set of educational programs which have equal power.

Linear Growth

To describe the general logic of this approach, we first need to conceptualize the notion of the "power" of an educational program. This power is a quantitative property, so we can imagine two educational programs which have equal power. Let us suppose that we have two such programs, and we call them A and B. Suppose also that program A deals with students in grade 3, and program B deals with students in grade 4. Then, if we consider a particular student, and he exerts the same level of effort for grade 3 and for grade 4, we would expect that, by definition,

his growth during grade 3 would be the same as during grade 4. This is what is meant by saying that it must be possible to compare a student's growth during one time interval with his growth during another time interval. Note that, from the point of view of the scale we desire, this implies that growth in the scale should be linear over time for any given individual.

More concretely, if Johnny Smith gains 10 units between September of grade 3 and June of grade 3, and if we can safely assume that the program in grade 4 is equal in power to that in grade 3, then we can expect him to grow 10 units between September of grade 4 and June of grade 4. If this same equality of program power is assumed for all the grades, then the trace of Johnny Smith's level of knowledge from grade 1 to grade 6 would be linear. It would appear as in Figure 1A.

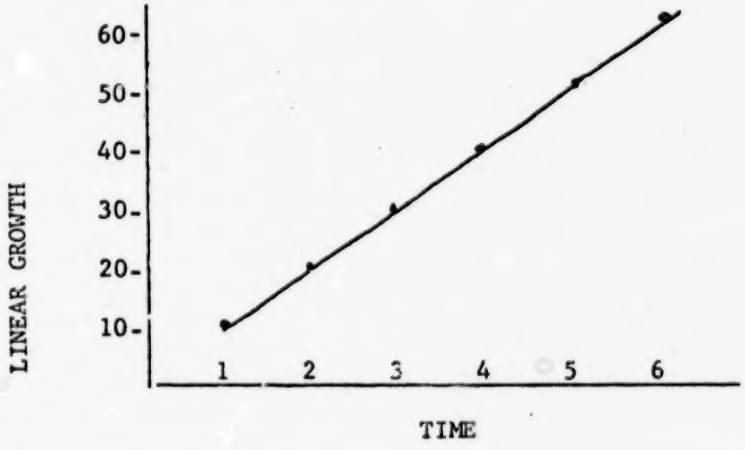

Figure 1A

Unlike the pattern just described, the growth shown by the M.A.T. standard scores or the S.R.A. growth scores is generally like that of Figure 1B. With this kind of growth pattern (which presumably does re-

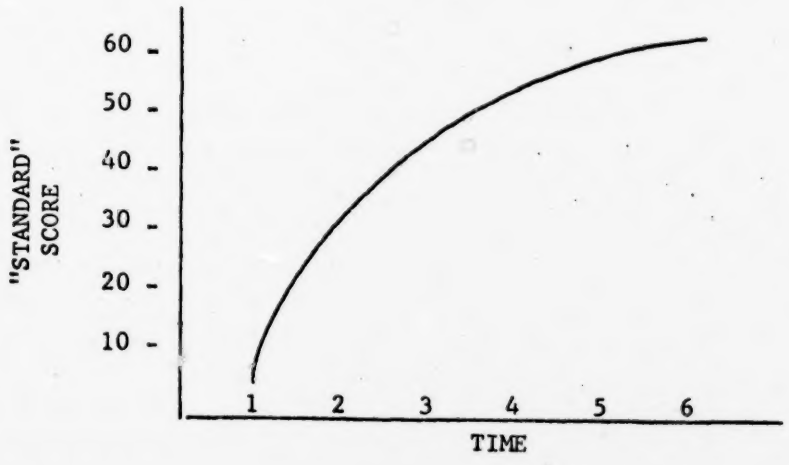

Figure 1B

flect the psychometrically true pattern of increase in the trait) there is no direct way to compare growth between grade 4 and 5 with growth between grade 1 and 2. Thus, these scores are not appropriate for comparing the results of two different programs except under very unusual conditions.

It should be pointed out that these "unusual" conditions occur when the initial scores and the learning rates of the children in program A are exactly the same as those in program B, and when there is also no differential regression caused by differential matching. To achieve these conditions is basically to achieve the classical experimental design, in which allocation of individual students to program A or B is random. Thus, the point emerges that one way (though probably not often practical, if recent past experience is a guide) to circumvent the whole dilemma of score format is to use strict randomization of assignment.

It may have occurred to the reader that the scores described above and shown in Figure 1A are not unfamiliar. They are, in fact, exactly the grade-equivalent score already used for other purposes. It is a defining property of the ordinary grade-equivalent score that, for the reference group on which the scores are based, these scores show a perfectly linear growth rate over time. This is an important point and makes the grade-equivalent score a strong candidate for use in program analysis.

Fan-Spread

Unfortunately, however, the grade-equivalent score lacks another property which would be desired in a score used for analysis. To see what that desirable property would be, imagine again that we have two educational programs whose "power" we know to be equal. Suppose however, that this time both of the programs are designed for use with children of the third grade. Suppose also that we apply program A to a group of children whose initial achievement level is 2.7, and we apply program B to a group of children whose initial achievement level is 3.2. For the sake of simplicity, we can even assume that there is only one child exposed to each program, or that in each of the groups, every child has exactly the same initial score as all the others in his program. Thus, we sidestep any arguments about the distribution of scores in each class. We would find that the gains shown by the children in program B (in which the initial level was higher) were greater than the gains of the children in program A. This difference in observed gain would not be due to any difference in program power, but instead to the fact that rate of gain in

score is almost invariably found to be positively correlated with initial score when grade-equivalent scores are used.

Grade-equivalents as a score format have been castigated (Dyer, 1971) because they possess this "fan-spread" property. That is, the graphic presentation of grade-equivalent scores over time shows that initially disadvantaged children (or more accurately, those who initially have a low score level) fall progressively further behind each year (See Figure 2).

A number of educators have rejected this pattern, and likewise the score format which produces it, because it suggests that the school is denying its impact to those who clearly need it most; it seems to be helping the "rich get richer and the poor get poorer." Our position is that this indictment is unjustified and that, even if it were justified, that would be no reason for throwing away the score form.

The absurdity of this rejection can be seen by comparing the situation revealed by grade-equivalents with the fact that in any long-distance foot race, the faster runners gradually move farther and farther ahead (measured in feet, meters, or inches) as time progresses. Yet, no one suggests that our scales of distance be rejected. Rather, what is done is to classify racers into approximately equal-speed groups, and then compare their performances. To the extent that the level of knowledge reached by a child after a length of time depends on his effective learning potential, to that extent the differences in growth rate indicated by grade-equivalents are real, but are totally irrelevant to the comparison and analysis of educational programs.

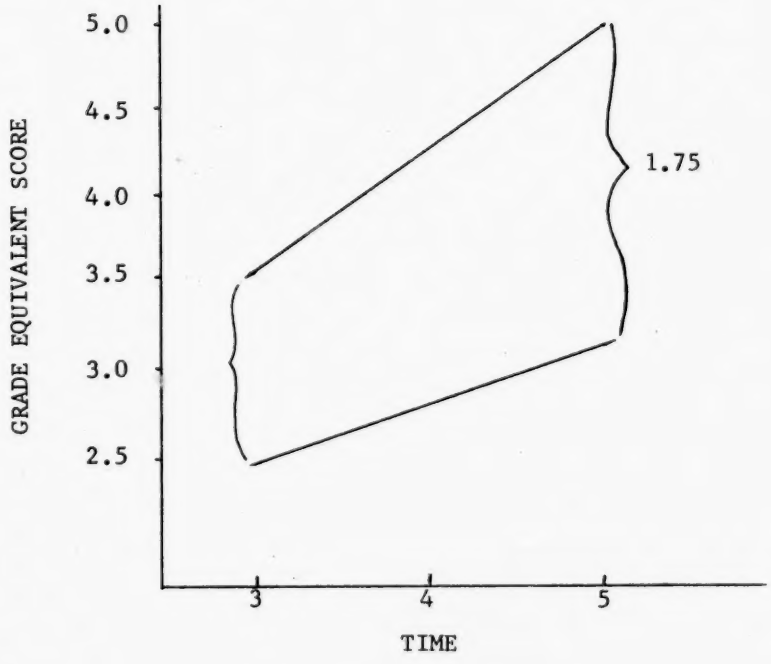

Figure 2

Using Grade-Equivalents Properly

Before going on to discuss some possible ways to deal with the fact that growth rate is proportional to initial level (because initial level serves as an imperfect but good indicator of effective learning potential), we need to mention one class of practical situations in which it does not represent a problem. In general, there is no reason to think that the effective learning potential of a particular student will change between time interval 1 and time interval 2. Naturally, concrete evidence would make this assumption questionable or even untenable; but in the absence of contradictory evidence, it can stand. Thus, regardless of what a particular student's individual learning rate is, his growth in terms of grade-equivalents during time interval 1 can be compared with his growth during time interval 2. Any differences that are observed under these conditions are probably the result of differential power of the two programs. The same reasoning holds if we have a cohort of students, and compare the average of their individual growths during the second time interval with the average of their individual growths during the first time interval.

It seems likely that this longitudinal comparison of single children, or intact groups of children, is what Lindquist and Hieronymus (1964) had in mind when they recommended the grade-equivalent scores as appropriate for measuring growth. Since they were writing in the more traditional context where the program is regarded as a constant and the question concerns the rate of development of the individual child, it is reasonable to guess that they were assuming that the program to which the child was exposed was the same at the two times. Under that assumption, differences

between the growth observed during the first time interval and that observed during the second would be cause for counseling the child or perhaps complimenting him on an outstanding effort.

Note too that this discussion brings out the interdependence of assumptions here. If the growth in the first time interval is not the same as that during the second time interval, it is possible only to say <u>something</u> has changed. Whether that change originates in the learning rate of the child (or children) or in differential power of the two programs is a question that must be settled by examination of the relative plausibility of these two explanations. In this connection, too, it might be wise to recall the cautions noted by Campbell and Stanley (1967) as to the possible distorting effect of history on a design of this general sort.

The practical conclusion that has emerged from the discussion thus far is simply the following. For situations in which there is measurement on the same children at three or more time points (at least two time intervals), it is legitimate to compare the growth (measured as differences in grade-equivalents) of individual children who have experienced the two programs as a means for comparing the programs. This strategy will become increasingly important as we accumulate more and more files of good-quality longitudinal data on achievement tests.

<u>Sub-Groups by Growth Rate</u>

The next part of our discussion considers the question of an appropriate score format when there is no way to use the child as his own control or when strict randomization has not been used. This is the most common situation. The approach we suggest for dealing with it is as straightforward (in principle, at least) as that used in the athletic

world. We would simply categorize students into groups which are internally homogenous as far as effective learning rate is concerned, and then make our comparisons only within these groups. This is, logically, parallel to the typical categories or handicapping systems used in most sports (boxing's weight classes, auto racing's classes, golf's handicaps, etc.).

The difficulty arises when we try to make this classification in practice. It is in this area that we most need empirical work and dissemination of results to provide a general pool of benchmark information for all researchers. Some first steps in this direction are apparent. For example, in several recent reports by researchers dealing with programs for improving the performance of disadvantaged children (Donaldson, 1971; U. S. Office of Education, 1972), there are statements to the effect that the "normal" or "expected" gain rate for disadvantaged children is about 0.7 grade-equivalent units per year. Thus, there has been an informal and crude partitioning of the gain rates into two categories (ordinary, and disadvantaged). The major problem with this particular classification scheme is that it is still extremely crude.

In fact, it is cruder than it need be for most studies. During the past year, this author was involved as a consultant and analyst on a research project dealing with disadvantaged children. The project, sponsored by the U. S. Office of Education, examined the feasibility and impact of offering monetary incentives to teachers and parents to improve school effectiveness. In the course of that project, we needed to calculate "expected" gains for each child in the project in order to determine whether individual teachers would receive cash bonuses at the end of the year. The field workers on this study needed to calculate

a set of expected gains which would be perceived as fair, not just by researchers in an academic context, but by real teachers for whom the gains would imply payment or lack of it.

Because the project dealt only with schools containing a large proportion of severely disadvantaged children, it might be thought that the typical gain rate of 0.7 would be a reasonable number for use in all classes. However, in a number of the schools and grades, the students were ability-grouped, which meant that one teacher might teach only the relatively slow students in that school while another teacher might teach only the relatively bright children. Researchers and teachers quickly and independently arrived at the conclusion that a more refined and specific set of expected gains was needed.

To provide these more specific benchmarks, the project workers used an approximation that seemed the best available under the circumstances. This approximation was developed by calculating the cohort-to-cohort difference between each adjacent grade in each school, separately for the upper third, middle third, and lower third, of all students in the school at that grade (Planar Corporation, 1972). Clearly, this procedure involves some assumptions that can be questioned. On the other hand, it is equally clear that it provides a substantial improvement in precision of prediction as compared to using a single number such as 0.7. As a matter of fact, the calculated gains to be expected ranged from about 0.2 to about 1.3 grade-equivalents per class; and, for those students in the middle third of the ranking, were usually not far from the 0.7 used in the other studies.

This solution, developed hastily under the pressure of real deadlines,

stands as a sensible compromise among a variety of conflicting criteria. The particular characteristics of the solution adopted in the Incentives Project are less important than the kinds of thought processes it reflects. In that project, there was a practical need to obtain expected gain estimates that would be as precise as possible, yet these estimates had to be provided within narrow constraints of time and money. In this situation, and in view of the absence of available gains data on similar populations to provide distributional information, there was reliance on a direct approach -- using cross-sectional data on the project schools as a substitute for the actual gains data. Evidently, the approximations were adequate, as is indicated by the fact that the subsequent post-test gains actually obtained tended to be distributed fairly closely around the predicted values, and were not systematically higher or lower for teachers regardless of the ability level of students they taught.

SUMMARY AND CONCLUSIONS

This paper has argued that a good deal of the recent controversy surrounding the uses of standardized tests is unnecessary; indeed, this controversy distracts attention from other related problems on which work is needed and possible. The wide interest in measures of achievement arises primarily because several recent large-scale evaluations of educational programs have made use of standardized achievement tests as if they provided interval scales for the variables of interest. In fact, most of the scales commonly used with standardized tests were not designed as interval scales. There are, however, some special scales offered by publishers with the claim that they have interval properties.

These latter scales are justified only by a fairly weak argument, that is, by the appeal to a normality of distributions which may not be the actual situation. More importantly, the scales so proposed, even if they are accepted as being what they claim, are demonstrably not interval scales for the variable which is of interest in program evaluation, namely, program impact. Only a scale that yields equal changes when any child is exposed to a program with a fixed "power" can meet this criterion.

Of the generally familiar scales, the one that seems most adaptable for this purpose is the grade-equivalent scale. This scale has been mistakenly attacked in recent years, because it seems to show patterns that some persons find threatening. The fact is, however, that this scale does have one of the two desirable properties needed in any scale for program analysis -- it yields linear growth for an individual child. Therefore, in situations where there is comparison of the same person's reaction to two programs, or where strict randomized assignment has been used, the grade-equivalent scale is perfectly appropriate as a basis for calculating gains.

For those more frequent situations in which there is a comparison of two non-equivalent groups a problem called "fan-spread" enters the picture. One very sensible and practical procedure that meets the problem of "fan-spread" is to stratify the population under examination into a number of subgroups according to the best available indicator of their effective learning potential. In many cases, the best available indicator will be the initial score, but other kinds of data can be used as substitutes or supplements to the initial scores. Once these subgroups have been defined, their expected rate of change can be estimated by

looking at the observed rate of change for similar groups under (presumably) similar conditions.

The practicality of this approach was demonstrated in the recent Incentives in Education project. For that project, cross-sectional data based on the same schools and the same students were used to create the benchmark. However, there is a wide variety of possibilities for creating these benchmarks. Thus, this approach, analogous to the calibration of a physical scale, provides a practical and flexible way to develop the kinds of scales that we need to analyze educational programs adequately.

There is no short-cut, general solution to the problem of developing benchmarks for a variety of situations, but there are direct and feasible ways in which progress can be made. One useful activity would be to compile tabulations of the distributions of observed gains in achievement test scores under various conditions. There are a number of data files from which such tabulations could be made without enormous effort. The material for this sort of tabulation exists not only in the files of several large scale research projects, but also in the files of several large school districts which administer standardized tests routinely. Work of this sort is underway now at Hopkins and elsewhere. As results from this sort of work accumulate, individual investigators will be less confined by the limitations of their own data and their own budgets in setting benchmarks.

As already mentioned, there are a variety of specific approaches which might be considered in developing the benchmark gains for the stratified grade-equivalent scales of program impact. One specific approach is illustrated in the Yardstick system (Pinkham, 1970), but

others are possible as well. Future research in this area will provide information about the advantages and disadvantages of different approaches and the relationship between their results. This work too is feasible and important, but in some cases will require the collection of richer data than is presently available.

References

Harvey A. Averch, Stephen J. Carroll, Theodore S. Donaldson, Herbert J. Kiesling, John Pincus, "How Effective is Schooling? A Critical Review and Synthesis of Research Findings," (Santa Monica, Rand Corporation R-596, April 1972)

Battelle Laboratories, Final Report on the Office of Economic Opportunity Experiment in Educational Performance Contracting, (Columbus, Ohio, Battelle Columbus Laboratories, March 1972)

Peter M. Blau and W. Richard Scott, Formal Organizations, (San Francisco, Chandler, 1962)

Samuel S. Bowles and Henry M. Levin, "The Determinants of Scholastic Achievement: An Appraisal of Some Recent Findings," Journal of Human Resources, vol. 3, #1, Winter, 1968)

Glen G. Cain and Harold W. Watts, "Problems in Making Policy Inferences from the Coleman Report," American Sociological Review, vol. 35, #2, April 1970, pp. 228-242.

Donald T. Campbell, "Temporal Changes in Treatment-Effect Correlations: A Quasi-Experimental Model for Institutional Records and Longitudinal Studies," Proceedings of the 1970 Invitational Conference on Testing Problems, (Princeton, N.J., Educational Testing Service, 1971)

Donald T. Campbell and Albert Erlebacher, "How Regression Artifacts in Quasi-Experimental Evaluation Can Mistakenly Make Compensatory Education Look Harmful," in Jerome Hellmuth, editor, The Disadvantaged Child, Volume 3, Compensatory Education, A National Debate, (New York, Brunner/Mazel, Inc., 1970)

Donald T. Campbell and Julian C. Stanley, Experimental and Quasi-Experimental Designs for Research, (Chicago, Rand McNally, 1967)

David K. Cohen, "Politics and Research: The Evaluation of Large Scale Social Action Programs in Education," Review of Educational Research, vol. 40, #2, 1970, pp. 213-238.

Harold L. Cohen and James Filipczak, A New Learning Environment, (San Francisco, Jossey-Bass, 1971)

James S. Coleman, "Incentives in American Education," Report #40, Center for Social Organization of Schools, The Johns Hopkins University, February 1969

James S. Coleman and Nancy L. Karweit, Information Systems and Performance Measures in Schools, (Englewood Cliffs, N. J., Educational Technology Publications, 1972)

James S. Coleman, Introduction to Mathematical Sociology, (New York, Free Press, 1964)

James S. Coleman, Ernest Campbell, et al., Equality of Educational Opportunity, (Washington, D. C., U. S. Office of Education, July 1966)

Lee J. Cronbach and Lita J. Furby, "How Should We Measure Change - Or Should We?", Psychological Bulletin, vol. 74, #1, 1970, pp. 68-80.

Theodore S. Donaldson, Data Requirements for Evaluation: A Review of Educational Research, (Santa Monica, Calif., Rand Corp. R-932/LACS, December 1971)

Henry S. Dyer, "The Role of Evaluation in Accountability," Proceedings of the Conference on Educational Accountability, Chicago, (Princeton, N. J., Educational Testing Service, June 1971)

Robert L. Ebel, "The Social Consequences of Educational Testing," School and Society, vol. 92, 1964, pp. 331-334

James Fennessey, "Some Problems and Possibilities in Policy-Related Social Research;" Social Science Research, vol. 1, #4, December 1972, pp. 359-383

John C. Flanagan, "Units, Scores, and Norms" in E. F. Lindquist, editor, Educational Measurement, (Washington, D. C., American Council on Education, 1951)

John C. Flanagan, The Cooperative Achievement Tests: A Bulletin Reporting the Basic Principles and Procedures Used in the Development of their System of Scaled Scores, (New York, American Council on Education, 1939)

Eric F. Gardner, "Comments on Selected Scaling Techniques with a Description of a New Type of Scale," Journal of Clinical Psychology, vol. 6, 1950, pp. 38-43

Eric F. Gardner, Determination of Units of Measurement which are Consistent with Inter and Intra Grade Differences in Ability, (unpublished doctoral dissertation, Harvard University, Graduate School of Education, 1947)

James W. Guthrie, "A Survey of School Effectiveness Studies," in Do Teachers Make a Difference?, (Washington, D. C., U. S. Office of Education, 1970)

Chester R. Harris, editor, Problems in Measuring Change, (Madison, Wis., University of Wisconsin Press, 1967)

Howard Kirschenbaum, Rodney Napier, Sidney B. Simon, Wad-Ja-Get, (New York, Hart Publishing, 1971)

Stephen Klein, "The Uses and Limitations of Standardized Tests in Meeting the Demands for Accountability," (Center for the Study of Evaluation, University of Los Angeles, Evaluation Comment, January 1971, vol. 2, #4)

Roger Lennon, "Accountability and Performance Contracting," invited address to the American Educational Research Association Convention, New York, February 1971

E. F. Lindquist and A. N. Hieronymus, Manual for Administrators, Supervisors, and Counselors, Iowa Tests of Basic Skills, (Boston, Houghton Mifflin, 1964

Melvin W. Lifson, Decision and Risk Analysis for Practicing Engineers, (Boston, Cahners Books. 1972)

Frederic M. Lord, "A Paradox in the Interpretation of Group Comparisons," Psychological Bulletin, vol. 68, #5, 1967, pp. 304-305

Frederic M. Lord, "Statistical Adjustments When Comparing Preexisting Groups," Psychological Bulletin, vol. 72, #5, 1969, pp. 336-337

Quinn McNemar, The Revision of the Stanford-Binet Scale, (Boston, Houston-Mifflin, 1942)

Walter N. Durost, et al., 1970 Metropolitan Achievement Tests, (New York, Harcourt Brace Jovanovich, 1970)

David W. Miller and Martin K. Starr, The Structure of Human Decisions, (Englewood Cliffs, N. J., Prentice Hall, 1967)

Robert J. Nash and Russell M. Agne, "The Ethos of Accountability: A Critique," Teachers College Record, vol. 73, February 1972, pp. 357-370

David B. Orr, "MAT Scaling, Procedures and Comments," unpublished memo to The Planar Corporation, March 1972

Robert G. Owens, Organizational Behavior in Schools, (Englewood Cliffs, N. J., Prentice Hall, 1970)

Fred O. Pinkham, "The Yardstick Project: Management Science On Line in the Schools," Educational and Urban Society, vol. 3, #1, 1970, pp. 71-98

Planar Corporation, Incentives in Education Project, Impact Evaluation Report, (Washington, D. C., October 1972)

James Popham, as quoted in Education Daily, May 25, 1972, p. 5

Alice M. Rivlin, Systematic Thinking for Social Action, (Washington, D.C., Brookings Institution, 1971)

Peter H. Rossi and Walter Williams, editors, Evaluating Social Programs, (New York, Seminar Press, 1972)

Robert E. Stake, "Objectives, Priorities, and Other Judgemental Data," Review of Educational Research, vol. 40, 1970, pp. 181-212

Robert E. Stake, in R. Tyler, R. Gagne, M. Scriven, editors, "Toward a Technology for an Evaluation of Educational Programs", Perspectives of Curriculum Evaluation, (Chicago, Rand McNally, 1967)

Science Research Associates, Evaluating Educational Growth, (Chicago, 1969)

Marshall S. Smith and Joan S. Bussell, "Report Analysis: The Impact of Headstart" Harvard Educational Review, vol. 40, 1970, #1, February, pp. 51-104

S. S. Stevens, "A Metric for the Social Consensus," Science, vol. 151, February 1966, pp. 530-540

L. L. Thurstone, "A Method of Scaling Psychological and Educational Tests," Journal of Educational Psychology, vol. 16, 1925, pp. 433-451

U. S. Office of Education, The Effectiveness of Compensatory Education: Summary and Review of the Evidence, (Washington, D. C., 1972)

Jonathan R. Warren, "College Grading Practices: An Overview," (Princeton, N. J., Educational Testing Service, 1971)

Charles E. Werts and Robert L. Linn, "Lord's Paradox: A Generic Problem," Psychological Bulletin, vol. 72, #6, 1969, pp. 423-425

Charles E. Werts and Robert L. Linn, "A General Linear Model for Studying Growth," Psychological Bulletin, 1970, vol. 73, #1, pp. 17-22

Westinghouse Learning Corporation - Ohio University, The Impact of Head Start, (Report to the Office of Economic Opportunity, June 1969)